CUPCAKES WERE SCREAMING: neutral machine translation

Marielle Egert, Lisa Yankovskaya

Introduction

The main idea of this project is to analyse translations of baseline model and remove the identified shortcomings.

The baseline model is OpenNMT-py [1] with default parameters and BPE [2] segmentation (30000).

Analysis

We analysed baseline translation and found the following shortcomings:

1. Content words are missed (ex. 4).
2. The part of sentence is missed after “dot” in the middle of sentence (ex. 1).
3. Incorrect translations of words (ex. 2, 3).
4. New words are added (ex. 5).

Our steps

- The most easiest problem is the second one. To solve it we replaced all dots by “_PUNCT_”.
- To solve the third problem we decided to replace BPE by WordPiece [3, 4] segmentation and vary beam size.
- To solve the first and the fourth issues we used transformer model.

Our models

We used sockeye [5] as a sequence-to-sequence framework for Neural Machine Translation and transformer [6] as an encoder-decoder architecture. Also, we changed the number of layers from two to six. We built two models with BPE (70000) and WordPiece (50000). We tried several beam sizes: 5, 8, 10 and 12. The best results for both models are obtained when the beam size is ten.

Difficulties

- Incorrect human translation: we found several sentences where human translators add new information to translation (ex.4) or remove some information from translation.
- Analysis of sentences was difficult due to lack of knowledge of Estonian.

Examples

- **Example 1**
  Source: 4. Otsustamine, millal midagi vaadata ja mida vaadata.
  Human: 4. Deciding when to see something, and what to see.
  Baseline: Four.
  Model with BPE: Deciding when to look at something and look at.
  Model with WP: 4. Decide when to look at something and watch what.
  Our comments: Our approach to replace dot by another symbol works.
- **Example 2**
  Source: Ungaris leiti, et peaaegu 96% lampidest on ohtlikud.
  Human: In Hungary, nearly 96% of the lights were found to be hazardous.
  Baseline: In Hungary, almost 96% of sheep were found to be dangerous.
  Model with BPE: In Hungary, almost 96% of sheep were found to be dangerous.
  Model with WP: In Hungary, almost 96% of the lamps were found to be dangerous.
  Our comments: The model with WordPiece showed more interpretable result than the model with BPE. However, it did not lead to higher BLEU.
- **Example 3**:
  Source: Esiteks vaatab see osakesi mitte kui punktikesi ...
  Human: First, it views particles not as points ...
  Baseline: First, it looks at the particulates ...
  Baseline with another beam size: Firstly, it looks at the particles ...
  Our comments: As we can see, the same model but with another beam size (ten) gave us the correct word: “particles” instead of “particulates”.
- **Example 4**
  Source: Oktoobril avalikustab EL ka läbi aegade esimese Euroopa arenguaruande.
  Human: On 22 October, the EU will also publish the first-ever European development report.
  Baseline: The EU will also make public the first European progress report on time.
  Model with BPE: On October, the EU will also make the first European development report through time.
  Model with WP: On October, the EU will also disclose the first European Progress Report of old times.
  Our comments: “October” is missed in the baseline model, both our models corrected this shortcoming.
- **Example 5**
  Source: Pädevuste enesehindamisse on kaasatud kogu personal.
  Human: When making the self assessment of competences the whole staff is involved.
  Baseline: The self-evaluation of the competences is the same as that of the European Union.
  Model with BPE: All personnel are involved in self-evaluation of competences.
  Model with WP: The entire staff shall be included in the jurisdiction.
  Our comments: The baseline model has a phantom expression “European Union”, both our models corrected this shortcoming.

Conclusion

Due to lack of time we did not compared the models with the same size of vocabulary and the same number of layers but that have different architecture: RNN vs Transformer. It is worth noting, the transformer model with six layers and a larger vocabulary was trained in less than two days, whereas RNN model (OpenNMT-py) with two layers and a smaller vocabulary was trained about 3.5 days.

BLEU score was higher for model with BPE, but results of the model with WordPiece in some cases look better (example 2). Likely, the increase in the number of word segmentation for WordPiece would give us an increase in BLEU value.

Besides that, it would be interesting to try some grammar checker for translations. Some of translations have grammar mistakes, for example, “uncorrect” instead of “incorrect” or incorrect prepositions, like “on September” instead of “in September”.

References

1. https://github.com/OpenNMT/OpenNMT-py
2. R. Sennrich, B. Haddow and A. Birch (2016): Neural Machine Translation of Rare Words with Subword Units
3. Y. Wu, M. Schuster, Z. Chen, etc (2016): Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. (https://arxiv.org/abs/1609.08144)
4. https://github.com/google/sentencepiece
5. https://github.com/awslabs/sockeye
6. A. Vaswani, N. Shazeer, N. Parmar, etc (2017): Attention is all you need (https://arxiv.org/abs/1706.03762)

lisa.yankovskaya@gmail.com, marielle@egert-hb.de